PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-272686

(43)Date of publication of application : 08.10.1999

(51)Int.Cl. G06F 17/30
G06F 17/27

(21)Application number : 10-070688

(22)Date of filing : 19.03.1998

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>

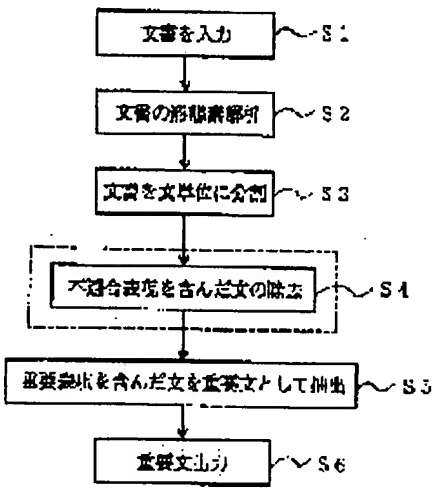
(72)Inventor : HORII MUNEYUKI
MATSUOKA KOJI
TAKAGI SHINICHIRO

(54) METHOD AND DEVICE FOR IMPORTANT DOCUMENT SENTENCE EXTRACTION AND RECORD MEDIUM WHERE IMPORTANT DOCUMENT SENTENCE EXTRACTING PROGRAM IS RECORDED

(57)Abstract:

PROBLEM TO BE SOLVED: To easily extract an important sentence from a document with high precision.

SOLUTION: An improper expression table wherein improper expressions are described and an important expression table wherein important expressions are described are prepared; and a morpheme analysis (S2) of an inputted document is taken, the analyzed document is divided (S3) into sentences, and sentences including improper expressions are removed (S4) from the document divided into the sentences by referring to the improper expression table. From the document from which the sentences including the improper expression have been removed, sentences including important sentences are extracted (S5) as important sentence by referring to the important expression table. Here, the process for removing the sentences including the improper expressions is omitted in some cases and the sentences including the important expression may be extracted as important sentence directly from the document divided into the sentences by referring to the important expression table.



LEGAL STATUS

[Date of request for examination] 10.05.2001

[Date of sending the examiner's decision of rejection] 07.09.2004

[Kind of final disposal of application other than the

examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] The document important sentence extract approach characterized by extracting a sentence including an important expression as an important sentence out of the document which inputted the document, carried out morphological analysis of the this inputted document, divided this document by which morphological analysis was carried out per sentence, and was divided into this sentence unit with reference to the important expression table which described the important expression.
[Claim 2] The document important sentence extract approach characterized by setting the document after removing the sentence which included the unsuitable expression out of the document in the document important sentence extract approach according to claim 1 with reference to the unsuitable expression table which described the expression unsuitable as an important sentence as the object of an important sentence extract.
[Claim 3] The document important sentence extractor characterized by to have a means extract a sentence including an important expression as an important sentence out of the document divided into said sentence unit, with reference to a means input a document, the means which carry out morphological analysis of said inputted document, a means divide said document by which morphological analysis was carried out per sentence, the important expression table which described an important expression, and said important expression table.
[Claim 4] A means to input a document, and the means which carries out morphological analysis of said inputted document, A means to divide said document by which morphological analysis was carried out per sentence, and the unsuitable expression table which described the expression unsuitable as an important sentence, The means which removes the sentence which included the unsuitable expression out of the document divided into said sentence unit with reference to said unsuitable expression table, The document important sentence extractor characterized by having a means to extract a sentence including an important expression as an important sentence out of the document which removed the sentence including said unsuitable expression, with reference to the important expression table which described the important expression, and said important expression table.
[Claim 5] The treatment process which carries out morphological analysis of the document which is the record medium which recorded the document important sentence extract program for extracting an important sentence from a document, and in which computer reading is possible, and was inputted, The unsuitable expression table which described the expression unsuitable as an important sentence to be the treatment process which divides the document by which morphological analysis was carried out per sentence is referred to. The treatment process which removes the sentence which included the unsuitable expression out of the document divided per sentence, The record medium characterized by having the treatment process which extracts a sentence including an important expression as an important sentence out of the document which removed the sentence including an unsuitable expression with reference to the important expression table which described the important expression.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention] This invention relates to the document processing system by the computer, and relates to the document important sentence extract approach and document important sentence extractor which extract the important sentence of a document for the purpose of the epitome of a document etc., and the record medium which recorded the document important sentence extract program for it in detail.
[0002]
[Description of the Prior Art] In order to survey efficiently a lot of information, especially the perusal information on WWW (World Wide Web) and the electronic information of an electronic mail and to acquire only required information, the epitome technique of a document is indispensable.
[0003] Conventionally, in order to summarize a document in the document processing system by the computer, considering an important sentence as a technique of extracting automatic, "the technique using the word frequency in a document" and the "technique of analyzing document structure" are well used out of a document. "The technique using the word frequency in a document" is the technique of choosing the sentence in which the ranking ranking-contains a high word for each word which appears in a document based on the frequency of occurrence of a word unit, and the frequency of occurrence of a document unit as an important sentence by carrying out statistics processing of a lot of documents. On the other hand, "the technique of analyzing document structure" is the technique of choosing an important sentence as an index for the language expression with which the block which should have the theme from the structure of a document is identified, an important sentence is chosen from the inside, or installation of subject, conversion, etc. are expressed.
[0004]
[Problem(s) to be Solved by the Invention] Although the "technique using the word frequency in a document" of the conventional technique turns into an effective approach when a lot of document data can be prepared on the other hand -- (1) -- (2) which needs a lot of document data -- a word (generally the frequency of occurrence in a document is high) with not necessarily high ranking When the frequency of occurrence of (3) relevance words by which the word with little frequency of occurrence in other documents is not necessarily contained in the important sentence is very high, an applicable word is contained in almost all sentences, and there is a trouble of many sentences being chosen.
[0005] Moreover, "the technique of analyzing document structure" is effective if document structure is well analyzable, but document structure must be able to be analyzed correctly and it speaks about it, and when aimed at a document with difficult analysis, and the document in which it does not have finite structure like an electronic mail like language, there is a problem which does not arrive to an important sentence extract.
[0006] This invention was made in view of the trouble of the above-mentioned conventional technique, and aims at offering the document important sentence extract approach of extracting the sentence which is a precision easy and high out of a document, and included the important expression as an important sentence, a document important sentence extractor, and the record medium that recorded the document important sentence extract program for it.
[0007]
[Means for Solving the Problem] In order to attain the above-mentioned purpose, this invention prepares the table which described an expression and an important expression unsuitable as an important sentence, and is characterized by extracting a sentence including an important expression as an important sentence for the document after removing the sentence which included the unsuitable expression out of the document. In addition, the processing which removes the sentence which included the unsuitable expression depending on the case may be omitted.
[0008]
[Embodiment of the Invention] Drawing 1 is the processing flow Fig. of the document important sentence extract

approach of this invention. As shown in drawing 1 , a document is inputted by this invention approach (step S1). Carry out morphological analysis of the inputted this document (step S2), divide this document by which morphological analysis was carried out per sentence (step S3), and the unsuitable expression table which described the expression unsuitable as an important sentence first is referred to. The important expression table which removed the sentence which included the unsuitable expression out of the document divided per sentence (step S4), next described the important expression is referred to. The sentence which included the important expression out of the document which removed the sentence including an unsuitable expression is extracted as an important sentence (step S5), and the this extracted important sentence is outputted (step S6). Here, the processing which removes a sentence including the unsuitable expression of step S4 may be omitted depending on the case, and the sentence which included the important expression directly with reference to the important expression table may be extracted as an important sentence out of the document divided per sentence.

[0009] Drawing 2 is the outline block diagram of one example of the document important sentence extractor of this invention. A this writing important sentence extractor consists of storage 40 which stores the body 20 of document processing system equipment which has each processing facility of the document input unit 10, the morphological analysis section 21, the simple sentence division section 22, the unsuitable sentence removal section 23, and the important sentence extract section 24, the document output unit 30 and the word dictionary 41, and the unsuitable expression important expression table 42 and 43. This configuration is realized by the so-called computer system. Here, the document input device 10 is the generic name of the communication interface for a document input with drivers, such as keyboard, image scanner, and FD- and CD-ROM, or WWW, or an electronic mail. The body 20 of document processing system equipment is the so-called CPU body which has an internal memory for storing a program, necessary data, etc. to perform. The document output units 30 are a display and a printer. A store 40 is a hard disk etc.

[0010] In drawing 2 , the document input unit 10 inputs the document set as the object of an important sentence extract, and sends out the inputted document to the morphological analysis section 21. The morphological analysis section 21 performs morphological analysis for the document received from the document input unit 10 with reference to the word dictionary 41, and sends out this document by which morphological analysis was carried out to the simple sentence division section 22. The simple sentence division section 22 divides the document received from the morphological analysis section 22 per sentence. It is the same as usual so far. In the simple sentence division section 22, the document divided into this sentence unit is sent out to the unsuitable sentence removal section 40.

[0011] The unsuitable sentence removal section 23 removes the sentence which included the unsuitable expression out of the document received from the simple sentence division section 22 by referring to the unsuitable expression table 42. The expression unsuitable as an important sentence is described as morphological information and sentence positional information by the unsuitable expression table 42. All the sentences that be described on the unsuitable expression table 42 and that shifted and included that expression are removed out of a document by the unsuitable sentence removal section 23. The document which removed the sentence including an unsuitable expression is sent out to the important sentence extract section 24.

[0012] The important sentence extract section 24 extracts a sentence including an important expression by referring to the important expression table 43 out of the document received from the unsuitable sentence removal section 23. The important expression is described as morphological information and sentence positional information by the important expression table 43. All the sentences that be described on the important expression table 43 and that shifted and included that expression are extracted out of a document as an important sentence by the important sentence extract section 24. The extracted important sentence is outputted to the document output unit 30.

[0013] Next, the technique of extracting an important sentence from the document by this invention is explained using an example.

[0014] With the document input unit 10, the case where the following electronic mail documents are inputted is considered.

<Input> Yokosuka information communication link He is Takeishi of Takagi Mr. Mita business affairs. I have always been indebted.

It is the wish of a system urgent meeting.

The inventory control system which your company could supply is downed this morning, and is working by the current emergency mode. The main situations are as follows.

- An initial screen is not displayed even if it accesses from a terminal.
- Backup data are unrestorable.

Trouble is in business with an end-of-the-year stage, and it must work on countermeasures immediately. Can't a telephone be obtained urgently? Since an immediate steps meeting is held from 15:00 of today in the Mita business-affairs head

office conference room at 17:00, I want to ask you for attendance. Although it feels sorry, I need your help well.
[0015] In the morphological analysis section 21, morphological analysis of the above-mentioned input-statement document is carried out with reference to the word dictionary 41. For example, the result of having carried out morphological analysis of "being Takeishi of the Mita business affairs" among the above-mentioned input-statement documents becomes like drawing 3 . In drawing 3 , the text and a dividing point, and the middle express the standard notation of each word, and the lower berth expresses [an upper case] the part of speech of each word. Morphological analysis of all the documents inputted with the document input unit 10 is carried out like drawing 3 by the morphological analysis section 21.
[0016] In the simple sentence division section 22, the input-statement document by which morphological analysis was carried out is divided per sentence. The above-mentioned input-statement document is divided per sentence as follows. However, below, a statement number is given and shown in order of the divided sentence on account of explanation. Moreover, morphological information is omitted.
1 The Yokosuka information communication link Mr. Takagi 2 He is Takeishi of the Mita business affairs.
3 I have always been indebted.
4 It is the wish of a system urgent meeting.
5 The inventory control system which your company could supply is downed this morning, and is working by the current emergency mode.
6 The main situations are as follows.
7 An initial screen is not displayed even if it accesses from - terminal.
8 - backup data are unrestorable.
9 Trouble is in business with an end-of-the-year stage, and it must work on countermeasures immediately.
10 Can't Telephone be Obtained Urgently?
11 Since an immediate steps meeting is held from 15:00 of today in the Mita business-affairs head office conference room at 17:00, I want to ask you for attendance.
12 Although it feels sorry, I need your help well.
[0017] In the unsuitable sentence removal section 23, when it investigates whether either of the expressions described by the unsuitable expression table 42 is contained to all the sentences divided in the simple sentence division section 22 and at least one is contained, the sentence is removed from the document received from the simple sentence division section 22 as a sentence including an unsuitable expression, i.e., a sentence which cannot turn into an important sentence.
[0018] The example of the unsuitable expression table 42 is shown in drawing 4 . The unsuitable expression table 42 consists of sentence positional information and morphological information. Sentence positional information has described the sentence location used as the object which confirms whether the expression described using morphological information is included. Usually, it is "*" and is aimed at all sentences. For example, only a top sentence will be applicable if it is a "head sentence." Morphological information consists of tetrad of 1 or more sets of initial entries, a surface notation, a standard notation, and a part of speech. An initial entry specifies the location of the following morpheme. It means that "next" has the following morpheme immediately after, and means that it may be separated from "far" of the following morpheme. Moreover, it expresses that "end" is the sentence end and "-" means that there is next no morphological information. It means that "*" in a surface notation, a standard notation, and a part of speech matches all.
[0019] It sets in each sentence by which division was carried out [above-mentioned], for example, is 1. The Yokosuka information communication link In Mr. Takagi, the part of "appearance (sentence end)" matches the expression of the table ID 100 of the unsuitable expression table 42. Therefore, it becomes the sentence removed from a document. Similarly, it is 2. He is Takeishi of the Mita business affairs.
** and the part of "being Takeishi" -- a table ID 101 -- a match and 3 I have always been indebted.
** and the part of "being **** to care" are a match and 12 to a table ID 102. Although it feels sorry, I need your help well.
** -- "-- well -- thank you for your consideration -- " -- a part matches a table ID 103. in addition, the statement number 12 -- "-- between " and "wishes" -- "-- well -- " -- although it enters -- a table ID 103 -- "-- since the initial entry of " is "far", the "wish" may be separated.
[0020] Thus, in the unsuitable sentence removal section 23, four sentences of statement numbers 1, 2, 3, and 12 are removed from an input-statement in the letter, and the following documents are sent out to the important sentence extract section 23.
4 It is the wish of a system urgent meeting.
5 The inventory control system which your company could supply is downed this morning, and is working by the current emergency mode.

6 The main situations are as follows.
7 An initial screen is not displayed even if it accesses from - terminal.
8 - backup data are unrestorable.
9 Trouble is in business with an end-of-the-year stage, and it must work on countermeasures immediately.
10 Can't Telephone be Obtained Urgently?
11 Since an immediate steps meeting is held from 15:00 of today in the Mita business-affairs head office conference room at 17:00, I want to ask you for attendance.
[0021] In the important sentence extract section 24, when it investigates whether either of the expressions described by the important expression table 43 is contained to all the sentences in the document with which the sentence including an unsuitable expression received from the unsuitable sentence removal section 2340 was removed and at least one is contained, it extracts as an important sentence.
[0022] The example of the important expression table 43 is shown in drawing 5 at drawing 4 . The configuration of the important expression table 43 is the same as the configuration of the unsuitable expression table 42 of drawing 4 . Therefore, in the important sentence extract section 24, it collates by the same technique as the unsuitable sentence removal section 23.
[0023] Four among the documents received from the unsuitable sentence removal section 23 to the above-mentioned example It is the wish of a system urgent meeting.
the part of ** "it is a wish" -- a table ID 202 -- a match and 10 Can't a telephone be obtained urgently?
** -- "-- it is -- ? -- " -- a part -- a table ID 200 -- a match and 11 Since an immediate steps meeting is held from 15:00 of today in the Mita business-affairs head office conference room at 17:00, I want to ask you for attendance.
** -- "-- thank you for your consideration -- " -- a part matches a table ID 201. In addition, the morphological information on a table ID 204 is 9. Trouble is in business with an end-of-the-year stage, and it must work on countermeasures immediately.
Although the part of ** "**" is matched, since it is in sentence positional information with a "head sentence", the statement number 9 which is not a head sentence does not correspond. (The head sentence in this case is a statement number 4) .
[0024] the above -- the three sentences above-mentioned in the important sentence extract section 24, 4 [i.e.,], It is the wish of a system urgent meeting.
10 Can't Telephone be Obtained Urgently?
11 Since an immediate steps meeting is held from 15:00 of today in the Mita business-affairs head office conference room at 17:00, I want to ask you for attendance.
It will be extracted as a ** important sentence.
[0025] In the example of drawing 2 , although the important sentence is extracted in the important sentence extract section 24 for the document which removed the sentence which included the unsuitable expression in the unsuitable sentence removal section 23, processing in the unsuitable sentence removal section 23 can be omitted, and an important sentence extract can also be performed, without removing an unsuitable sentence. In that case, since all the sentences of statement numbers 1-12 serve as an input to the important sentence extract section 50 in the above-mentioned example of a document, it is 12 in addition to three sentences of statement numbers 4, 10, and 11. Although it feels sorry, I need your help well.
It ******. ("-- thank you for your consideration -- " -- a table ID 201 -- match) .
[0026] Therefore, the important sentence extracted is 4. It is the wish of a system urgent meeting.
10 Can't Telephone be Obtained Urgently?
11 Since an immediate steps meeting is held from 15:00 of today in the Mita business-affairs head office conference room at 17:00, I want to ask you for attendance.
12 Although it feels sorry, I need your help well. It becomes four **.
[0027] Thus, although the statement number 12 which is not so important will also be extracted by omitting processing of the unsuitable sentence removal section 23, it is satisfactory at extent from which precision falls somewhat in the above-mentioned example of a document.
[0028] Next, the case where the following electronic mail documents are inputted is considered with the document input unit 10 as other examples.
He is <input> Tanaka.
"The 3rd information communication link study meeting" is held on April 25. This theme is "an agent communication link." In Mr. Sato's group, I think that it is the deep theme of relation. If there is a participating candidate, even Tanaka needs to inform. much participation -- I am waiting.

[0029] It passes through the morphological analysis section 21 and the simple sentence division section 22, and is 1. He is Tanaka.
2 Hold "the 3rd information communication link study meeting" on April 25.
3 This theme is "an agent communication link."
4 In Mr. Sato's group, I think that it is the deep theme of relation.
5 If there is a participating candidate, even Tanaka needs to inform.
6 much participation -- I am waiting.
It is sent out to the ** unsuitable sentence removal section 23.
[0030] It is 1 by referring to the unsuitable expression table 42 of drawing 4 in the unsuitable sentence removal section 23. He is Tanaka.
It removes ** picking. (A table ID 101 matches saying "He is Tanaka") .
[0031] Therefore, in the important sentence extract section 24, it is 2. "The 3rd information communication link study meeting" is held on April 25.
3 This theme is "an agent communication link."
4 In Mr. Sato's group, I think that it is the deep theme of relation.
5 If there is a participating candidate, even Tanaka needs to inform.
6 much participation -- I am waiting. It ******.
[0032] It is 2 by referring to the important expression table 43 of drawing 5 in the important sentence extract section 24. "The 3rd information communication link study meeting" is held on April 25.
5 If there is a participating candidate, even Tanaka needs to inform.
It is extracted by the match of ********** "**" (head sentence) and a table ID 204, and the match of "please contact me" and a table ID 203 as an important sentence.
[0033] When processing of the unsuitable sentence removal section 23 is omitted in this example, a statement number 2 is not extracted in the important sentence extract section 24. This is because a statement number 2 is no longer a head sentence, so sentence positional information of a table ID 204 is not fulfilled. Therefore, the important sentence extracted is 5. If there is a participating candidate, even Tanaka needs to inform.
A next door and precision will fall considerably.
[0034] In the above, the example of this invention was explained. Here, the processing flow of the document important sentence extract approach of this invention shown in drawing 1 may be recorded and sold to record media, such as FD or CD-ROM, as a document important sentence extract program in the format in which computer reading is possible. If the document important sentence extract program recorded on this record medium is installed in a computer, use with the operation gestalt of drawing 2 will be attained.
[0035]
[Effect of the Invention] According to this invention, the important sentence in a document can be easily extracted in a high precision as mentioned above only by preparing the table which described an expression and an important expression unsuitable as an important sentence. Especially this invention is effective in an electronic mail document etc.

[Translation done.]